Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/88/20241660

Virtual shadow puppet generation based on AlphaPose

Bangbang Sun

School of Software Engineering, Shandong University, Shandong, China

202222300279@mail.sdu.edu.cn

Abstract. This study aims to develop a novel motion capture method utilizing the AlphaPose project to support the production and analysis of shadow puppetry. Through precise capture and analysis of human body postures, we seek to enhance the performance quality of shadow puppetry, introducing modern technology to this ancient Chinese intangible cultural heritage art form, thereby facilitating its inheritance and development. This paper provides an introduction to the background and significance of the research, explores existing issues and shortcomings, and provides a detailed description of our methodology and the improvements achieved.

Keywords: Alphapose, shadow puppet, whole-body pose estimation, pose tracking.

1. Introduction

Originating in the Western Han dynasty, flourishing during the Tang dynasty, and reaching its peak in the Qing dynasty, shadow puppetry is a traditional Chinese performing art form that involves the manipulation of cut-out figures made of animal hide or cardboard, illuminated by light to tell stories. Despite being officially recognized as an intangible cultural heritage in 2011, this cultural treasure, with over two thousand years of heritage, faces numerous challenges in the face of the information age and the emergence of new forms of entertainment.

Challenges include the intricate craftsmanship involved in puppet creation, reliance on manual string manipulation or finger sensors for artistic presentation, high training costs, limited controllable movements of the puppets, and complex manipulation procedures. In recent years, the rapid development of multimedia has introduced virtual shadow puppets, offering new perspectives for the inheritance and development of traditional Chinese shadow puppetry. Virtual shadow puppets simplify the performance process, align better with contemporary artistic appreciation habits, and, when improving traditional stage puppetry, better showcase the artistic charm of shadow puppetry.

However, existing virtual shadow puppets often use Flash skeletons to connect various body components, employ physics engines for physical modeling, associate the model with various parts of Flash components, and control external Flash components by manipulating the physical model. Despite some improvements involving motion capture through gesture recognition using sensors, these methods still require performers to manipulate puppet movements manually, resulting in high training costs and imprecise control, making shadow puppetry less accessible to beginners and yielding suboptimal puppetry outcomes.

AlphaPose, while maintaining precision, achieves rapid pose recognition, approximately three times faster than traditional methods such as Mask-RCNN, for both single and multiple individuals. Furthermore, AlphaPose extends beyond body pose estimation to excel in capturing poses for facial

DOI: 10.54254/2755-2721/88/20241660

expressions and limbs. Leveraging AlphaPose's pose recognition and establishing a skeleton model, performers can drive shadow puppet animations simply by performing actions, eliminating the traditional training cost associated with manual manipulation and enabling anyone to quickly create shadow puppets.

Therefore, this paper proposes a virtual shadow puppet generation method based on AlphaPose, utilizing AlphaPose technology to capture human actions and driving shadow puppet movements through topological structure transfer to generate shadow puppetry.

To be more specific, our work contributes in the following ways:

- Utilizing AlphaPose with a top-down approach, employing the RMPE framework, and employing technologies such as SSTN, DPG, and p-NMS to address the human body pose estimation problem in different scenarios.
- Simplifying the human body skeleton model after obtaining the skeletal structure, making it similar to the shadow puppet model by discarding components like hand bone models.
- Estimating and transforming parameters such as rotation angles for the human body skeleton model to adapt it to shadow puppet movement poses.

2. Related Work

In previous work, Toshev and Szegedy employed deep neural networks for single-person keypoint detection, utilizing DeepPose to infer the positions of human body keypoints. They formalized the pose estimation problem as a regression task and employed end-to-end learning to directly predict the coordinates of keypoints from input images [1]. "Convolutional Pose Machines" proposed the segmentation of pose estimation tasks into multiple stages, progressively refining each stage to enhance the accuracy of pose estimation [2]. Newell et al. introduced a stacked network architecture called "Hourglass," performing feature extraction and keypoint regression at multiple scales [3]. In the paper "Deeply Learned Compositional Models for Human Pose Estimation" [4], a hierarchical compositional architecture was introduced, leveraging deep convolutional neural networks to model each part of the human body, reducing computational complexity and improving performance. In "Human Pose Estimation with Spatial Contextual Information" [5], two simple yet effective modules, Cascade Prediction Fusion (CPF) network for predicting keypoints and Pose Graph Neural Network (PGNN) for correcting higher-level predicted keypoints, were proposed. Subsequently, "Toward fast and accurate human pose estimation via soft-gated skip connections" [6] improved residual blocks and introduced a mixed network.

For multi-person keypoint detection, Fang employed a top-down approach, addressing errors in human skeleton keypoint detection caused by cropping differences through a spatial transformation network [7]. The Face++ team adopted a strategy of first detecting simpler keypoints, followed by detecting more challenging keypoints, and finally detecting the most difficult or invisible keypoints [8]. Qi proposed a spatial shortcut network (SSN) specifically for pose estimation tasks, combining feature map movement and attention mechanisms in a module called the feature shifting module (FSM), facilitating the spatial flow of information [9].

Regarding the generation of shadow puppet animations, the majority decompose the shadow puppet model into multiple parts and assemble them according to different needs to quickly construct the shadow puppet model. For example, in the research and design of digital animation skeleton generation for Haining shadow puppetry [10], adding a skeleton to shadow puppet animations was proposed, constructing a skeletal network to enable full-range swinging of each joint.

DOI: 10.54254/2755-2721/88/20241660

3. Method

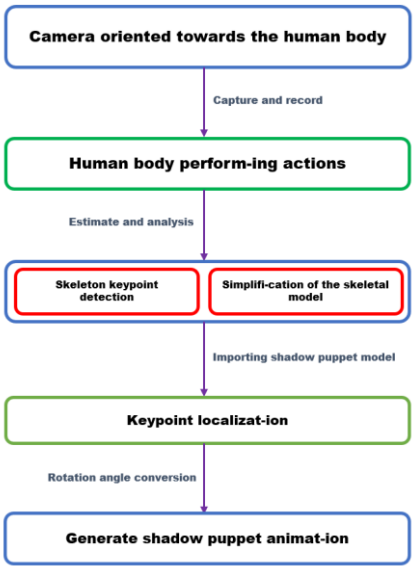

Figure 1. Firstly, we utilize a camera to capture and record the movements of the human body. After estimating and analyzing the results, we perform skeletal keypoint detection and simplify the skeletal model for the recorded body movements. Subsequently, we import the shadow puppet model and fit it to the recently simplified model, achieving the localization of shadow puppet skeletal keypoints. Finally, the rotational angles of the model's movements are translated into visual actions for the shadow puppet, thus realizing the generation of shadow puppet animation.

We utilize AlphaPose to extract skeletal nodes during the motion of the target. Subsequently, we simplify the skeletal node model and transform the rotational angles of the human body skeleton to fit the joint composition of the shadow puppet model. Finally, based on this, we create shadow puppet animation.

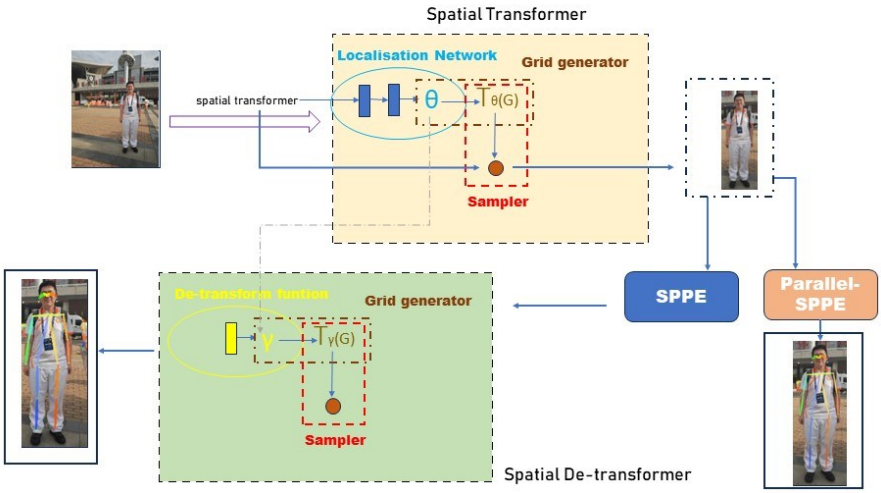

Figure 2. After importing an image, the parameters $\boldsymbol{\theta}$ generated by the localization network are used, and the inverse function $\boldsymbol{\gamma}$ is computed. Subsequently, a Grid generator + Sampler is employed to extract the region of the human body in the image. The model architecture of Parallel-SPPE is consistent with the main SPPE, but during the training process, the parameters in the model are frozen and do not participate in training. This branch serves as a regularization mechanism for the Spatial Transformer Network (STN).

DOI: 10.54254/2755-2721/88/20241660

3.1. AlphaPose

3.1.1. Symmetric STN Structure and Parallel SPPE

Since the SPPE algorithm is trained for single-person objects and is sensitive to localization errors, the human body region box obtained from object detection algorithms may not be very suitable for SPPE. The effectiveness of SPPE can be significantly improved through a pruning method. SSTN + Parallel SPPE can enhance the performance of SPPE effectively even under imperfect human body region detection results. The mathematical formula for using STN to extract high-quality human body region boxes is as follows:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = [\boldsymbol{\theta_1} \quad \boldsymbol{\theta_2} \quad \boldsymbol{\theta_3}] \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

where

$$\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3}$$

are vectors in two-dimensional space. And

$$\{x_i^s, y_i^s\}, \{x_i^t, y_i^t\}$$

are for the coordinates before and after transformation. At the end of SPPE, the pose results are mapped to the original bounding box of the human body. SDTN reflects this result back to the coordinates in the original image and requires calculation:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = [\boldsymbol{\gamma_1} \quad \boldsymbol{\gamma_2} \quad \boldsymbol{\gamma_3}] \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \tag{2}$$

The parameters in question have the same meaning as the parameters in STN.As SDTN is the reverse node of STN, the following expression can be obtained:

$$[\gamma_1 \quad \gamma_2] = [\theta_1 \quad \theta_2]^{-1} \tag{3}$$

$$\gamma_3 = -1 \times [\gamma_1 \quad \gamma_2]\theta_3 \tag{4}$$

In order to propagate backward in SDTN (We consider the cost function to be $J(W, b)$)

$$\frac{\partial J(W,b)}{\partial [\theta_1 \quad \theta_2]}$$

it can be expressed as:

$$\frac{\partial J(W,b)}{\partial [\theta_1 \quad \theta_2]} = \frac{\partial J(W,b)}{\partial [\gamma_1 \quad \gamma_2]} \times \frac{\partial [\gamma_1 \quad \gamma_2]}{\partial [\theta_1 \quad \theta_2]} + \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial [\gamma_1 \quad \gamma_2]} \times \frac{\partial [\gamma_1 \quad \gamma_2]}{\partial [\theta_1 \quad \theta_2]} \tag{5}$$

To further

$$\frac{\partial J(W,b)}{\partial \theta_3} = \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \theta_3} \tag{6}$$

This makes the inaccurate bounding box accurate after going through STN+SPPE+SDTN. As shown in Figure 1.

3.1.2. P-NMS

To avoid redundant pose detections, the P-NMS (Pose Non-Maximum Suppression) method is proposed. If the pose is represented as

DOI: 10.54254/2755-2721/88/20241660

$$P_i\{(k_i^1, c_i^1), \ldots, (k_i^m, c_i^m)\}$$

where k,c denote the position and confidence, we select the keypoint with the highest confidence as a reference. Points close to it are eliminated through an elimination criterion. This step is repeated until only one pose remains. The elimination criterion is defined as follows:

We define the pose distance matrix:

$$d(P_i, P_j \mid \Lambda)$$

to measure the similarity between poses. The elimination criterion is given by:

$$f(P_i, P_j \mid \Lambda, \eta) = \mathbb{I}[d(P_i, P_j \mid \Lambda, \lambda) \geqslant \eta] \quad (7)$$

Where η represents the elimination criterion, and Λ represents a set of parameters for the function $d(\cdot)$.If $d(\cdot)$ is smaller than η, then the output of $f(\cdot)$ is 1, indicating that P_i should be eliminated relative to P_j.

3.1.3. PGPG

As a two-stage pose estimation method, appropriate data augmentation is necessary. One heuristic approach is to directly use the pedestrian detection boxes generated during the training phase. However, pedestrian detection boxes can produce only one detection result for each pedestrian. By utilizing PGPG (Pose Guided Proposal Generator), the quantity of these results can be significantly increased, generating a large number of training samples and enhancing the system's capabilities.

3.2. Skeleton model transformation

AlphaPose typically provides information about multiple skeletal nodes, including various parts of the human body. However, digital shadow puppets usually focus on essential key points such as limbs (shoulders, elbows, wrists, hips, knees, ankles), head, and torso. Therefore, it is necessary to filter out the relevant nodes. Fortunately, AlphaPose itself has a default set of keypoint outputs, including Nose, Leye, Reye, Lear, Rear, Lshoulder, Rshoulder, LElbow, RElbow, LWrist, RWrist, LHip, RHip, LKnee, RKnee, LAnkle, Rankle. Therefore, we only need to retain: Lshoulder, Rshoulder, LElbow, RElbow, LWrist, RWrist, LHip, RHip, LKnee, RKnee, LAnkle, Rankle.

3.3. Rotation Angle Transformation

Since the final virtual shadow puppet relies on animation tools like flash for display, merely using coordinates is not sufficient for direct movement. Therefore, we need to convert the skeletal keypoint coordinates into rotation angles. Taking the connection between the left arm and torso as an example, let the coordinates of the initial Lshoulder in the first state be$(x1, y1)$, and the coordinates of the Lshoulder in the next state be $(x2, y2)$. The absolute rotation angle is then give

$$angle = atan2\left((x1 - x2), (y1 - y2)\right) \quad (8)$$

where

$$atan2(y, x) = \begin{cases} \arctan(y / x), & \text{if } x > 0 \\ \arctan(y / x) + \pi, & \text{if } y \geqslant 0, x < 0 \\ \arctan(y / x) - \pi, & \text{if } y < 0, x < 0 \\ \pi / 2, & \text{if } y > 0, x = 0 \\ -\pi / 2, & \text{if } y < 0, x = 0 \\ \text{undefined} & \text{if } x = 0, y = 0 \end{cases} \quad (9)$$

DOI: 10.54254/2755-2721/88/20241660

The same process applies to other body parts.

Due to the strong continuity of shadow puppetry movements, we can adjust the capture frequency and utilize automatic animation completion.

4. Experimental and Discussion

The computer configuration we are using is a 12th Gen Intel(R) Core(TM) i9-12900H 2.50 GHz CPU, 16GB RAM, with the Microsoft Windows 11 operating system. The generated visual result is as follows:

Figure 3. The generated outcomes are juxtaposed with reference images, presented vertically in the following order: the portrait images, the skeletal diagrams of human figures generated by AlphaPose, and the corresponding silhouette images thereof.

The evaluation metrics are as follows:

Table 1. evaluation metrics in this experiment

Number of Raw Key Points	Number of Key Points	Number of Renderings	Average FPS
17	10	20	8.74

From the experimental results, the transformation of various human body movements in the same direction as shadow puppetry is generally smooth. However, due to the inherent 2D nature of shadow puppetry models, if there is a change in the orientation of the human body or a significant alteration in angles, the shadow puppetry model may struggle to accurately replicate the movement. This issue can be addressed by later converting the shadow puppetry modeling into a three-dimensional representation, enabling more accurate handling of changes in orientation and angles.

DOI: 10.54254/2755-2721/88/20241660

5. Conclusion

In this project, we built upon AlphaPose to achieve the transfer of human body movements to shadow puppetry actions through skeleton redirection and rotation angle transformations. This allows the generation of corresponding shadow puppetry animations with just appropriate human body movements.

However, due to the substantial differences between shadow puppetry skeletons and human skeletons, authentic shadow puppetry animations involve many movements that are challenging for humans to imitate. This limitation results in relatively simple generated animations, and the orientation issue restricts the complexity of the actions. Additionally, variations in the form and attire of the shadow puppets can lead to deviations in the actions performed under the same human body movement, making it challenging to create a universal set of actions applicable to all shadow puppets.

In the future, digital shadow puppetry could be combined with technologies such as virtual reality (VR) to better optimize the transformation of shadow puppetry movements. This would enhance controllability, reduce performance difficulties, and facilitate the inheritance and development of this culturally rich and ancient form of intangible heritage among the younger generation.

References

[1] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1653-1660.

[2] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[3] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 483-499.

[4] Tang W, Yu P, Wu Y. Deeply learned compositional models for human pose estimation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 190-206.

[5] Zhang H, Ouyang H, Liu S, et al. Human pose estimation with spatial contextual information[J]. arXiv preprint arXiv:1901.01760, 2019.

[6] Bulat A, Kossaifi J, Tzimiropoulos G, et al. Toward fast and accurate human pose estimation via soft-gated skip connections[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 8-15.

[7] Fang H S, Xie S, Tai Y W, et al. Rmpe: Regional multi-person pose estimation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2334-2343.

[8] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7103-7112.

[9] Qi T, Bayramli B, Ali U, et al. Spatial shortcut network for human pose estimation[J]. arXiv preprint arXiv:1904.03141, 2019.

[10] Xu J. Research and Design of Digital Animation Skeleton Generation for Haining Shadow Puppetry[J]. Journal of Jilin Province Education Institute (Mid Issue), 2013, 29(11): 149-150.